AD-A044 556 FEDERAL AVIATION ADMINISTRATION WASHINGTON D C OFFICE--ETC F/G 5/9
OBJECTIVE METHODS FOR DEVELOPING INDICES OF PILOT WORKLOAD,(U)
JUL 77 W D CHILES

UNCLASSIFIED FAA-AM-77-15 NL

| OF |
AD
A044556

END
DATE
FILMED
10-77
DDC

(14) FAA-AM-77-15

(10) B S

(6) OBJECTIVE METHODS FOR DEVELOPING INDICES OF PILOT WORKLOAD

(10) W. Dean/Chiles
Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma

(11) July 1977 (12) 45 p.

Prepared for
U.S. DEPARTMENT OF TRANSPORTATION
Federal Aviation Administration
Office of Aviation Medicine
Washington, D.C. 20591

264320

hB

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-AM-77-15		

4. Title and Subtitle	5. Report Date
OBJECTIVE METHODS FOR DEVELOPING INDICES OF PILOT WORKLOAD	JULY 1977
	6. Performing Organization Code
	8. Performing Organization Report No.

7. Author(s)	
W. Dean Chiles	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aeromedical Institute P. O. Box 25082 Oklahoma City, Oklahoma 73125	11. Contract or Grant No.
	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address	14. Sponsoring Agency Code
Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, S.W. Washington, D.C. 20591	FAA

15. Supplementary Notes

Work was performed under Task AM-D-77-PSY-57.

16. Abstract

This paper discusses the various types of objective methodologies that either have been or have the potential of being applied to the general problem of the measurement of pilot workload as it occurs on relatively short missions or mission phases. Selected studies that have dealt with the workload measurement problem or some similar problem are reviewed in relation to their applicability to securing answers to operational questions. The types of methods are classified as: laboratory, analytic and synthetic, simulator, and in-flight. The paper concludes with a general discussion of the relative merits and some of the cautions to be observed in attempting to apply these methods and in trying to interpret the results with a view toward generalizing to operational situations.

17. Key Words	18. Distribution Statement
Performance Measurement Simulators Workload	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	43	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

OBJECTIVE METHODS FOR DEVELOPING INDICES OF PILOT WORKLOAD

I. Introduction.

The performance of the pilot as an aviation subsystem is conditioned by a large number of different factors, such as training, selection, and physical condition (on the personnel side) and the extent to which the principles of human engineering are applied to the design of the system and the mission profiles (on the hardware and operational side). For our purposes, we will assume that appropriate attention has been allocated to personnel factors. Thus, our primary focus will be on those aspects of human engineering (human factors, ergonomics) that relate to pilot workload as it may be affected by the overall design of the system.

The concept of workload is of special interest in that there is abundant evidence (at least of an anecdotal nature) that workload can be a "go/no go" modifier of the performance of the pilot as a functional subsystem, especially under emergency conditions. Therefore, finding or developing an appropriate methodology that yields reliable and valid measures of pilot workload is a goal that, if achieved, should lead to important gains in safety and mission accomplishment through the resultant system design and procedural modifications.

Our ultimate concern in the measurement of workload must be the determination of the manner and extent that workload affects the probability of mission success. Thus, in this context, it is appropriate to raise the traditional engineering questions related to the probability of "failure" of the pilot as a functional sybsystem. From the point of view of reliability engineering, we might say as a first approximation that an acceptable level of workload for a given phase of a mission would be characterized by a set of system-induced (system in its broadest sense) task demands such that the probability is equal to or greater than some specified value that the pilot will be able to satisfy those demands and successfully complete that mission phase without compromising subsequent mission phases. (Clearly, the probability value selected for one-time, high-priority missions, for multiple missions, and for routine operations would likely be different.)

The literature in this area is quite clear on one point. There is no generally accepted definition of the term "workload." Some authors would use the term primarily to refer to input loading; e.g., the number and nature of the displays (and controls) that must be used by the pilot in performing his job. Others would use the term to refer to how hard the pilot has to work; these authors tend to prefer biomedical and/or subjective indices of workload. Still other authors emphasize those aspects of workload that relate to performance; e.g., speed and accuracy of response.

A. A Working Definition of Workload. No attempt will be made to arrive at a formal, comprehensive definition of workload; the problems in developing such a definition are numerous and formidable. However, it seems necessary to offer some sort of working definition--even though it be rather nonspecific and largely descriptive--of the way the term will be used here before meaningful discussion of measurement methodology in the area can be undertaken. Therefore, for the purposes of this paper, level of pilot workload will be assumed to be an hypothetical concept that is determined by or (if you prefer) related to the aggregate of the task demands placed on the pilot by the system during some relatively short-duration mission or phase of a mission coupled with the actions required of the pilot to satisfy those task demands. The actions required may be overt or they may be covert. They may be physical, they may be mental, they may be perceptual, they may be oral, or they may be some combination of any or all of these. There may be purposes for which it is appropriate to talk about system demands independent of pilot actions in considering workload. However, in the present discourse it will be assumed that, to the extent a system demand is not followed by suitable and timely action on the part of the pilot, the mission phase will have been completed in less than an acceptable manner (if it is completed at all). In other words, demands that do not require action (either overt or covert) are not really demands; and actions that are initiated for reasons other than to satisfy a system demand (and are potentially disruptive of mission accomplishment) should be eliminated by training and operating procedures. Thus, "stimulus" and "response" will not be treated separately.

Although for purposes of exposition a general definition of workload is adopted, it should be clearly understood that the goals and intents of a given measurement effort are the important determiners of how workload should be defined and what methodology should be adopted for a specific application. For example, one designer/researcher may need to know simply which of two alternative--but otherwise satisfactory--single-purpose displays makes a smaller contribution to the pilot's workload. Another designer/researcher may need to know how quickly, if at all, the pilot can manually operate a device that is normally hydraulically or electrically powered. Numerous other differences in purposes and, hence--by implication--methodologies can be readily imagined. More will be said on this topic later, but it is not our intent to be dogmatic--especially about unsettled issues.

B. Outline. The remainder of this text will consist of six sections: Some Rudiments of Measurement Theory; Laboratory Methods; Analytic and Synthetic Methods; Simulation Methods; In-Flight Methods; and Discussion, Recommendations, Cautions, and Conclusions. The approach that will be used in the research-oriented sections will be to describe selected programs in which particular methodologies have been applied, and, where appropriate, data will be presented to give an indication of the kinds of results achieved. No attempt at a comprehensive review will be made; for reviews of relevant areas, the reader is directed to the following reports: Gerathewohl (25) has written a concise evaluation of the literature on

workload definition and measurement; Gartner and Murphy (23) survey and critique the literature on pilot workload and fatigue; White (55) reviews task analysis methods in relation to workload specification; Jahns (29) provides a general review and evaluation of the literature on operator workload; and Spyker, Stackhouse, Khalafalla, and McLane (50) review the workload literature in relation to quantitative, subjective and physiological methods.

II. Some Rudiments of Measurement Theory.

This section is not in any way intended to be a definitive exposition on measurement theory. However, certain basic concepts of measurement theory will come up in later sections and it seems expedient to mention and briefly explain them before proceeding. (Some readers may wish to skip this section.)

A. Validity. The first and perhaps most important notion to be dealt with is validity. Ultimately, this simply means, "Are we really measuring what we intend to be measuring?" The answer to this question, in the most precise use of the term, assumes the existence of a criterion. For example, in the field of selection, we might want to select only those aviation candidates who have a high probability of completing flight training; our criterion, then, would be successful completion of training (and perhaps final average grade). The validity of the selection measure would thus be determined by the accuracy with which it predicts which trainees will graduate. Unfortunately, in the workload area we have no such criteria, and, therefore, we must rely primarily on what is called "content validity"--which really amounts to expert, professional opinion. Still another kind of validity, "face validity," can be important in motivating test subjects; in this sense, (face) validity means the test situation appears to be like the job of the pilot. (No small part of the expense of building simulators is devoted to trying to achieve face validity.)

B. Reliability. Reliability has several meanings that are applicable in varying degrees to the problem of workload measurement. In one use, it refers to the engineering characteristics of the measurement system and relates to the repeatability of a measure or phenomenon; with a constant known input, what is the variability of the output? That is, how accurately can the output be predicted from the input? Reliability in this sense involves internal characteristics of the test device, and the term is used to reflect the sensitivity of a measurement procedure to temperature changes, drift characteristics of components, etc. A second, closely related use of the term "reliability" depends not only on the above characteristics of the test equipment but also on the human behavior being measured. For example, in even the most carefully controlled experimental situation, the response latency of the human subject to the onset of a light will show variation across trials and across individuals; the amount of such variation will depend on the behavior being measured. In this use of the term, an approximation of the reliability estimate can be obtained by observing the

extent to which a group of individuals shows the same rank ordering on each of two measurements of the phenomenon per individual. This is generally referred to as test-retest reliability. It should be noted that the apparent reliability (i.e., the size of the reliability coefficient) is dependent on both the true reliability of the test or equipment used and the existence of stable individual differences in the behavior being measured. Thus, with highly trained, highly selected, skilled operators, the variability for a given individual from trial to trial may be as great as the variability across individuals on a given trial. Under such conditions, the measured reliability could appear to be rather low even though the basic measures are quite stable. In any case, if meaningful comparisons are to be made concerning workload variations, some estimate of the stability and precision of the measures must be secured. Otherwise, there is no way to determine whether an obtained difference in a measure is properly interpreted as being real or as being a result of chance factors.

C. Sensitivity. In any evaluation of alternative system designs or system operating procedures, it is necessary to have some index of the sensitivity of the measures to the variables being manipulated. For example, simple reaction time to an attention-getting signal calling for a single response is quite stable even when there are large changes in presumably important variables. The same is true of many simple tracking tasks. Perhaps the main reason for this stability is the extreme adaptability of the human operator. If the operator is confronted with a task situation in which he can concentrate all of his resources on the performance of the task, then, at least for relatively short intervals, he can maintain his performance of single tasks amazingly well. Thus, for example, if altitude were a variable of interest and simple reaction time were the measure used, we would conclude that performance is not impaired until the pressure altitude is somewhat in excess of 5,000 meters. Thus, such simplistic approaches could lead to questionable conclusions. What all of this means is that it is sometimes necessary either to do preliminary research or to add variables to the main research simply to get an index of the sensitivity of the measurement procedure to relevant variables.

D. Magnitude of Effect. If two alternatives (displays, for example) are exactly equivalent in terms of cost, weight, size, etc., then any reliable (statistically significant) superiority of one alternative over the other is sufficient basis for choosing the better alternative. However, if there are important differences between the two in terms of cost, weight, etc., then it is necessary to establish not just the statistical significance of a difference (if there is one) but, especially if the more expensive one is the better, how much better it must be to make in fact a practical difference. Expert, professional judgment plays a major role here.

III. Laboratory Methods.

From the point of view of methodology, there are three characteristics of "laboratory" methods that make them highly desirable. First, for most

laboratory tasks, it is possible to exercise very precise control over the performance demands imposed on the operator. One can with relative ease control the number of tasks that are active, the rates at which signals are presented, and the timing of the signals on individual signal sources as well as across sources. Second, "exact" duplication of test procedures is readily achieved. Third, laboratory methods in general can provide the highest precision of measurement that one is likely to achieve in the realm of operator behavior. Fourth, depending on the level of complexity of the experimental task structure, high equipment reliability is possible at relatively modest costs, and, because physical safety is not involved, any lack of mechanical or electrical reliability is primarily just a source of inconvenience. In addition, tasks can be selected and structured so that good test-retest reliabilities are common. And fifth, it is generally not terribly difficult to establish the sensitivity of the task measures to variables of known operational importance and behavioral potency.

A. Background Research. Early in the history of the behavioral sciences, there was considerable interest in the area of mental load in what would now be called an information processing context. These early efforts were directed at an attempt to break down complex reaction time into its constituent components. To illustrate how this breakdown was approached, assume that the operator is confronted with a red light on the right of a display and a green light a few centimeters to its left. Assume further that two response buttons are conveniently located for the use of the right hand. The subject is instructed to depress the rightmost button if the red light comes on and the left button if the green light comes on. Thus, the subject must decide which light came on and which button is correct. Assume now a different procedure: a number of responses are recorded in which only the red light and the rightmost button are present and other responses when only the green light and the leftmost button are present. With this procedure, the subject only has to become aware that a light is on and respond. The notion is that the difference between the average response time to the single-light/single-button conditions and the two-light/two-button condition provides an estimate of the "mental" processing time in recognizing whether the red or the green light has been illuminated in the latter condition. This general procedure has been expanded and permutated in a variety of ways. The well-established result is that if N signals are uniquely coordinated to N possible responses, then:

$$\text{Reaction Time} = a + b*\log_2 N$$

where a is the y-axis intercept, b is the slope constant, and $\log_2 N$ is the measure of information "H." Thus, it is seen in this very elementary case that performance is a function of task demand or workload.

B. Timing--Speed and Load Stress. Another line of laboratory research has been concerned with the timing of response in a monitoring situation. The notion of timing in skilled performance was first introduced by

Sir Fredrick Bartlett (4). The concept was further refined by Conrad (18), who proposed to define timing (of responses) as "creating the most favorable temporal conditions for response." Conrad treated load in his studies as being a function of the number of signal sources and considered load stress to be produced by increasing that number beyond some value. He used the term speed stress to refer to excessive rates of presentation of signals from a given source (or number of sources). Conrad found that subjects tended to alter the point of response initiation in a manner apparently designed to even out, temporally, the sequence in which they were required to take action. In a later study, Conrad (19) gave subjects limited control over the average rate at which signals would appear; this control gave subjects the opportunity to slow down the signal rate so they could successfully respond to essentially concurrent signals on separate displays; on the average, subjects did better under this condition. These results are suggestive of the desirability of, wherever possible, adopting designs and operating procedures that permit latitude in the exact point at which events must be initiated by aircrew personnel.

Knowles, Garvey, and Newlin (34) investigated speed and load effects in a different context; they were interested in display-control compatibility relationships. The part of their experiment that is of particular interest here is the comparison of a 10x10 matrix of lights (associated with a 10x10 matrix of response buttons) and a 5x5 matrix of lights (associated with a 5x5 matrix of buttons). The rate of presentation of information (not signals) was equalized across the two conditions; the rates used were 1.75, 2.25, 2.75, and 3.0 bits/second. They found that the effect of load (display size) had a greater effect on error rate than did rate of presentation of signals. (See Table 1.) They also found, incidentally, that subjects could

Table 1. Mean Errors Per 100 Stimuli*

	Speed (bits/s)			
Matrix	1.75	2.25	2.75	3.0
Small (5x5)	2.5	3.6	4.1	7.1
Large (10x10)	3.6	10.8	13.1	15.8

*Adapted from Knowles, Garvey, and Newlin (34).

respond at an average rate of 0.45 signal per second without errors in a self-paced mode whereas when the task was forced-paced at that same rate, subjects made 36 percent errors.

C. Secondary Loading Tasks. One general, more direct approach to the study of workload in the laboratory has been through the use of secondary or loading tasks. Knowles (36) summarizes early work of this sort and provides the general rationale for the application of the technique to workload measurement in a part-task simulation context. Knowles (page 156) states that auxiliary tasks are used ". . . with the intention of finding out how much additional work the operator can undertake while still performing the primary task to meet system criteria.

"Secondary tasks are used because primary part-task performance measures, in and of themselves, seldom reflect operator-load. . . . they seldom tell the price paid in operator-effort in meeting (the system) criterion." Knowles goes on to describe an earlier study, Knowles and Rose (35), in which a simulated lunar landing task was being investigated. He says that in that study: "The loading scores were sensitive to differences in problem difficulty; they reflected increased ease in handling the control task as a function of practice; they revealed differences in workload between members of a two-man crew; and they showed that the particular control law under consideration was unsatisfactory because of the extreme buildup of operator load during the last few seconds of the landing. None of these results was available from system performance criteria; i.e., time, fuel, miss-distances." (Emphasis added.) The basic approach in this method is to compare the levels of performance achieved on the "loading" task when performed alone with the levels achieved when it is performed in combination with the primary task; this difference is said to provide an index of the workload imposed by the primary task.

Benson, Huddleston, and Rolfe (5) reported a study in which, among other things, they evaluated a one-dimensional tracking task by using two altitude displays; performance was measured with each display with and without a secondary light-acknowledgment task. They found a small, consistent superiority of a counter-pointer display over a counter-only display with the tracking-only condition. When the secondary task was added, they found significant decrements in tracking with both displays with a significant superiority of the counter-pointer over the counter-only display. The secondary task showed significant decrements when added to either tracking task; the differences between display conditions were fully compatible with the findings for the tracking task--namely, the display that showed the better performance in tracking showed the lesser effect on the performance of the secondary task. They interpret the decrements in the primary tracking task to pose serious questions as to "the essential feature of the subsidiary task situation; namely, that consistent primary task performance is possible in two task conditions." Benson et al. (5) instructed their subjects that they were to attend to the secondary task only when they could properly do both jobs together. They interpret their results to suggest that subjects may not be able to comply with such instructions and discuss at some length whether and how subjects might be able to perceive that their performance is being maintained on the primary task. They also suggest the

possibility that a continuous primary task may be more likely to suffer decrements than a discrete primary task. Depending on the frequency characteristics of the display disturbances and the time it takes the subject to perceive which light has been illuminated, it is quite reasonable to expect that, on a probabilistic basis, looking at and responding to their secondary task would encourage error accumulation on their primary task.

It should be noted that Benson et al. (5) concluded that "there is no doubt that the presence of a second task added to the value of the experiment. . . ." Thus, their discussion of the changes in the primary task is related primarily to "theoretical" expectations as to how the secondary task technique should operate in practice. It could be argued that their experiment actually demonstrated two important findings: (1) the counter-pointer display is better in that it resulted in better performance (numerically in the case of tracking only and statistically in the case of the two-task situation); and (2) the counter-only display is more sensitive to possible distraction or interference from other tasks.

The question can also be raised as to whether the subsidiary task technique necessarily relies on the subject's achieving parity of performance on the primary task between the one- and two-task conditions. Clearly, Benson et al. (5) demonstrated in their experiment that useful information can be obtained from the technique when this assumed state of affairs does not obtain. If we consider one of the empirically based reasons that Knowles pointed at in using the technique, it is frequently the apparent absence of an effect on single tasks of possibly important variables that suggests the possible value of using secondary operator loading tasks. Thus, it could be argued that so long as changes in the primary task and the secondary task are compatible (i.e., lead to the same conclusions), we should not be overly concerned about changes in the primary task--changes that may be valuable data in and of themselves.

Senders (49) says there are four assumptions that underlie the secondary loading task methodology: (1) The operator is a single-channel system. (2) The channel has a fixed capacity. (3) The capacity has a single metric by which any task can be measured. And (4) the constituents of workload are additive linearly, regardless of the sources of the load. These assumptions are required if channel capacity is to be given formal status as that term is used in information theory. However, in the practical application of the secondary loading task methodology, I believe the first and second assumptions stated by Senders are of major significance only under certain conditions--for example, when neither the primary task performance nor the loading task performance changes when the two are performed simultaneously. In that event, although we would have learned something interesting about the two tasks, we could not be sure whether the primary task represents a "no load" condition, the operator has employed a previously "unused" channel, the operator has simply "expanded" his (single) channel capacity, or, what is most likely, the time requirements of the two tasks are such that the

performance of neither interferes with that of the other. The possible absence of linear additivity places a heavy burden of responsibility on the choice of the loading task; clearly, the loading task must have properties in the "additivity domain" that warrant generalization to the kinds of system tasks that might be coupled with the primary task being investigated. By the same token, the metric implied by the secondary task must also be applicable to possible system task requirements.

Perhaps the safest interpretation of the changes in the secondary task would be that they serve as an index of the spare time that the operator has while performing the primary task at criterion levels. But even in this interpretation it is necessary to make some kind of assumption regarding the ease of back-and-forth transition (primarily in terms of time) between the primary task and the particular secondary task being used. Rolfe (42), who provides an excellent review and discussion of the secondary task method of measuring workload, closes with the following caution: "The final word, however, must be that the secondary task is no substitute for competent and comprehensive measurement of primary task performance. The technique should always be looked upon as a means of gathering additional information rather than an easy way of gathering primary information." This caution should not be taken lightly, even though the study of Knowles and Rose (35) showed secondary task measures to be sensitive to important factors not revealed by the primary task measures.

D. Cross-Adaptive Loading Tasks. Kelley and Wargo (31) take the position that consistent performance on the primary task is a must. They offer data from a demonstration experiment using two subjects in which decrements on primary and secondary tasks are apparently not compatible; conditions that were ranked, in order of merit, A, B, C on the primary task were ranked B, A, C by measures from a secondary task. Their primary task was a two-dimensional, two-display compensatory acceleration tracking task; the secondary task consisted of two identical "warning" lights, one above the other, located where subjects could see them by peripheral vision but had to look at them directly to determine which light had been illuminated; response to the lights was made with a thumb switch located on the tracking control stick. When the lights task was active, one of the lights, selected at random, would turn on 0.44 second after the subject extinguished the previous light. The primary task variable of interest was display gain, of which there were three levels. Three test conditions were used: primary task only, primary task plus the loading task with independent programing (straight subject pacing), and primary task plus "cross adaptive" programing of the loading task. In this latter case, as long as tracking error (vector root-mean-square (RMS)) remained below the criterion level, one of the lights would be turned on as noted above. If error exceeded the criterion level, the lights task would be deactivated until tracking error again was below criterion. It is important to note that Kelley and Wargo (31) instructed their subjects to perform both tasks " . . . as well as they could and not to neglect one for the other." Thus, the concepts of primary and secondary are

somewhat blurred; the experimenter, without informing the subjects, had arbitrarily decided which was which. The previously mentioned findings from Kelley and Wargo, in which the inferences from the primary and secondary task performances were not compatible, were taken from the condition involving tracking plus the subject-paced loading task. The compellingness of their results suffers from several problems. First, only two subjects were used. Second, the display gain variable was significant for the tracking-only condition. Third, the display gain variable was significant for the subject-paced loading-task condition for one subject though not for the other. And, fourth, a cleaner evaluation of the cross-adaptive approach to using loading tasks would have resulted if task priorities had been manipulated through instructions, or whatever. However, the approach, overall, looks interesting and further evaluation of its characteristics vis-a-vis traditional loading-task procedures would appear to be warranted.

E. Memory Scanning Tasks. Another variation on the secondary task technique has been described by O'Donnell (41). This procedure is "an adaptation of an item recognition technique first described by Sternberg" (51,52). The basic approach is that the operator is required to learn a set of positive stimuli (so-called because their appearance calls for a positive response). Members of the positive set, frequently letters of the alphabet, are presented one at a time; generally, on half of the trials the stimulus is a member of a negative set. On the appearance of a letter, the operator is instructed to respond as quickly as possible by depressing a "yes" key if the letter is a member of the positive set and a "no" key if it is a member of the negative set. Under appropriate conditions, a linear relation exists between the size of the positive set (typically 1 to 8) and reaction time. The psychological theory behind the use of this task is that average reaction time with a given number of stimuli in the positive set can be broken down into three parts: (1) stimulus encoding, (2) memory scan, and (3) response selection and execution. For a given set of conditions, the first and third parts are assumed to be constant, whereas the second part is interpreted to be a direct reflection of memory scan speed and/or memory load. Thus, changes in the y-intercept value are assumed to reflect changes in the perceptual and/or response aspects of the task. Changes in the slope of the curve are assumed to reflect changes in the rate at which memory is scanned and/or the amount of memory load involved. In other words, the y-intercept value serves the same function as a measure from a secondary loading task as described previously; the higher the intercept (i.e., the longer the average response time), the greater the assumed loading produced by the primary task. In addition, a change in the slope of the response-time curve might be interpretable as a reflection of the amount of memory load added by the primary task. The value of this task as a loading task in the usual sense has been borne out by the results of preliminary studies conducted thus far. However, the possibilities with respect to its providing a measure of memory load are still to be demonstrated. It should be noted that earlier results reported by Darley, Klatzky, and Atkinson (22) suggest that the addition of memory load not directly related to the item recognition task does not affect the slope of the reaction time curve.

F. Synthetic Work Tasks. Operator workload has also received attention in an area of laboratory research that is concerned with "synthetic work." The rationale for the development of synthetic work tasks has been described in detail elsewhere (13,14); however, for those readers to whom the notion is new, a brief description of the techniques and philosophy will be given here.

The point of departure of the synthetic work approach is a behavioral analysis of the performance requirements placed on the operator by some particular aviation system or by a class of such systems in general. Tasks are then selected against a criterion of content validity (i.e., tasks are selected because they measure functions judged by experts in the field to be important to aircrew operations) as well as a general criterion of face validity (i.e., the tasks are configured to be acceptable to target populations, such as pilots). Consumer acceptance of the tasks has always been good (13). The resultant hardware is designed so that the selected tasks can be presented in any combination desired and individual tasks can be varied along both time constraint and task difficulty parameters. The original goals of the program in which the particular system to be described was conceived were the evaluation of procedural (e.g., work schedules), environmental (e.g., altitude), and pharmacological (e.g., alcohol) variables as these factors might affect complex performance.

Within the context of the way these tasks were developed and have been used, the notion of workload is a relative concept. However, from the beginning it was assumed that it would be desirable, if not necessary, to vary the apparent workload imposed on the operator from very light to near overload; overload is defined, for this purpose, as decrements on all or most of the concurrently performed tasks, even in the absence of any external stressor. Thus, extensive data have been collected on a variety of task combinations that, on a rationally defensible basis, would be expected to correspond to different workloads.

The specific tasks used involve monitoring of lights and meters (providing measures of reaction time), mental arithmetic, pattern discrimination, elementary problem solving, and two-dimensional compensatory tracking. The task combinations used in a study by Hall, Passey, and Meighan (26), involving an earlier version of what is called the Multiple Task Performance Battery (Chiles et al., 13), are shown in Table 2. Note that two basic conditions were examined--monitoring tasks only and "full battery" as specified in Table 2. If it is assumed that the subjects tended to treat the monitoring tasks as secondary (loading) tasks, then the performance levels on those tasks can be considered to be an index of the workload imposed on the operator by the different combinations of the other tasks. Figure 1 shows the response latencies on a normalized scale for the responses to the offset of any one of five green lights located one at each corner and one in the middle of the test panel. Figure 2 shows response times in seconds for the detection of a shift in the average value of the "randomly"

Table 2. Performance Schedule*

	Monitoring Only								Complex							
Auditory Vigilance	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Warning Lights	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Meter Monitoring	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Mental Arithmetic									X	X						
Problem Solving (Group)										X	X	X	X			
Pattern Discrim.														X	X	
15-Minute Interval	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8

*Adapted from Hall, Passey, and Meighan (26).

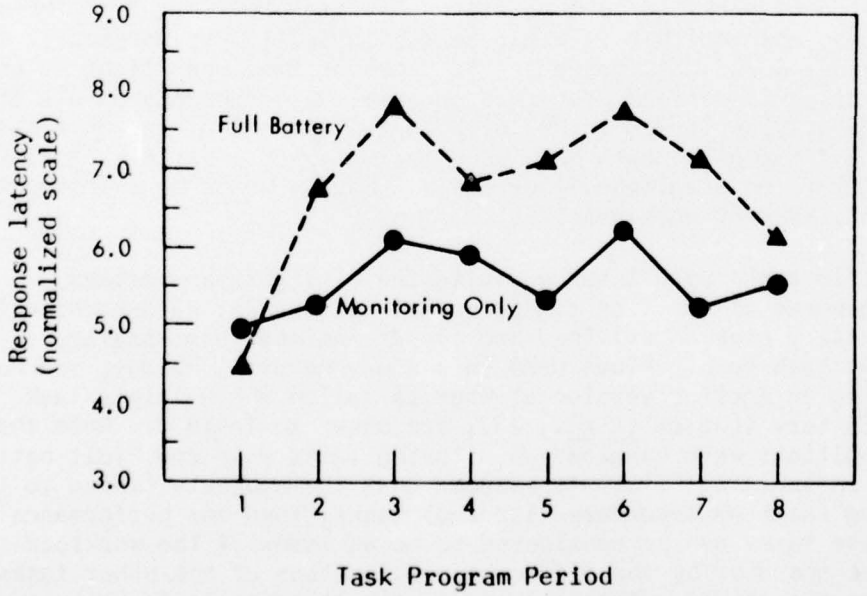

Figure 1. Mean response latency in detecting green warning-light signals during each 15-minute period of the basic 2-hour task program. (Adapted from Hall, Passey, and Meighan, 26).

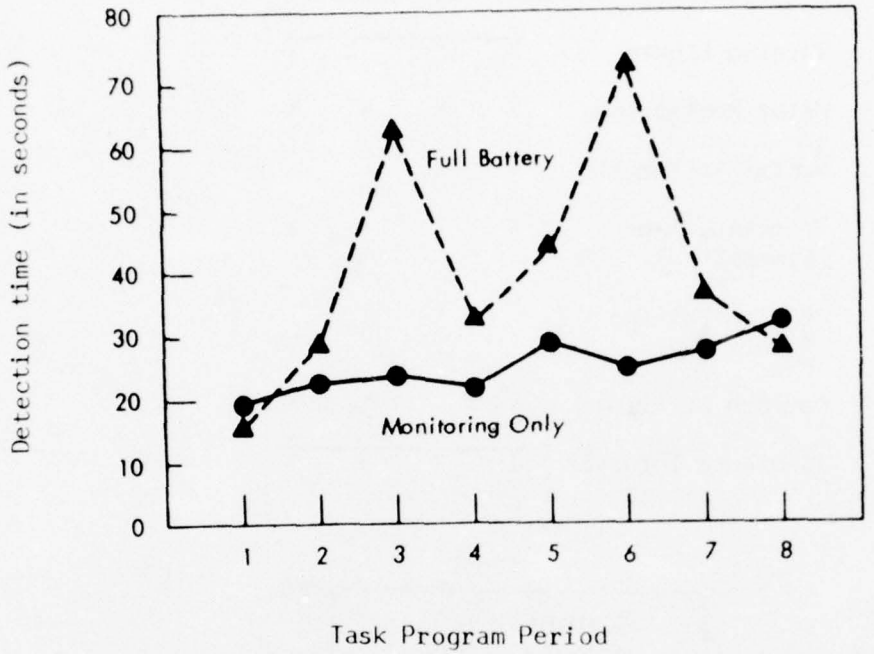

Figure 2. Mean detection time for correct detections of probability monitoring signals during each 15-minute period of the basic 2-hour task program. (See Table 2.)

wandering pointer of any one of four meters located across the top of the test panel. Each of these figures contains two curves--one for the given monitoring task performed with only the monitoring tasks active and one for monitoring performance as a function of the different "active task" combinations. Note that the first and the last points of the curves labeled "full battery" consist of only the monitoring tasks, thus providing "anchor points" for the curves. The normalizing scale applied to the data for the green-lights monitoring tends to suppress the apparent amplitude of the shift in response times, but the changes across task combinations are statistically significant. The changes in the meter-monitoring task are much larger and, of course, are also statistically significant.

The data shown in Figure 3 are from a later unpublished study using the task schedule shown in Table 3 and using pilots as the subjects. Figure 3 shows response times in seconds to the onset of red lights (physically paired with the green lights) and the offset of green lights. Figure 3 also shows the detection times in seconds for the meter-monitoring task. (Although the tasks are functionally the same as those used by Hall et al. (26), the data of these two figures were collected by using a new, computerized version of the Multiple Task Performance Battery.) For all three task measures, the differences across task combinations are significant. (It may or may not be

Table 3. One-Hour Task Schedule

Warning Lights	X	X	X	X
Meter Monitoring	X	X	X	X
Mental Arithmetic	X	X		
Tracking, Two-Dimensional	X			X
Problem Solving (Individual)		X	X	
Pattern Discrim.			X	X
15-Minute Interval	1	2	3	4

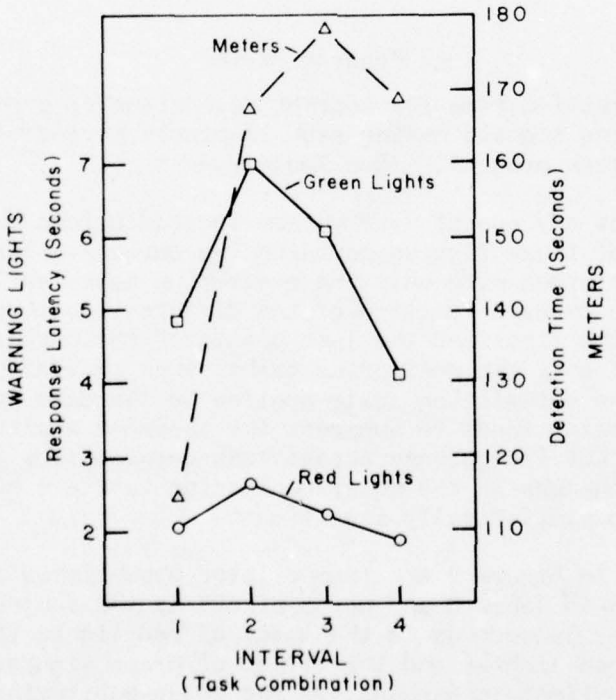

Figure 3. Monitoring performance as a function of task combination as shown in Table 3.

important that the longest response times for the light-monitoring tasks were associated with a different task combination than were the longest response times for the meter-monitoring task.) Significant differences were also found between task combinations for the tracking task (vector RMS error) and for the problem solving task (redundant responses). Neither the mental arithmetic task nor the pattern discrimination task showed significant differences as a function of task combination. This lack of differences could mean that these latter tasks are less sensitive to workload variations, or it could mean that they were given higher priorities by the subjects. Although a detailed evaluation of exactly how to account for the differences across tasks is not relevant to our purposes, some general observations are perhaps in order.

The data of Figure 3 are based on the mean of two 1-hour sessions; the subjects had had a total of about 7 hours of practice on the tasks before the first of these sessions and 10 hours of practice before the second. Among the literally hundreds of subjects who have learned to perform these tasks, it has been typical that the subjects initially have difficulty, for example, completing arithmetic problems in the allotted 20 seconds with any time to spare. Similarly, they frequently get "hung up" on the problem-solving task at the expense of the other tasks, even though they are reminded during training that they are to attend to all tasks. Thus, the learning procedure typically consists of first, acquiring skill on the individual tasks and, then, gradually learning to shift rapidly and efficiently from a given active task on which their attention may be focused at a given time to concurrent demands (e.g., the onset of a red light or another active task); or, on satisfaction of the momentary demands of the active tasks, they may shift to scanning the panel for monitoring signals. It is also clear that even at high levels of training, there are substantial individual differences in the smoothness and speed with which attention appears to be shifted from exercising one kind of behavioral process to another, different kind of process. For this and other reasons, a study was undertaken (30) to determine whether an independent (time sharing?) skill in this domain could be identified by using the techniques of factor analysis. In this study, the lights (red and green) and the meter-monitoring tasks were found to load on separate factors when performed as individual tasks. When performed as part of a complex task, these monitoring tasks all loaded on a third, independent factor. If these results, which suggest a possible time-sharing ability, should hold up on replication, important implications are suggested for the selection of subjects to be used in various kinds of tests of systems and system components.

The synthetic work methodology has yielded other results of relevance to the use of secondary loading tasks as measures of workload. In a study of the effects of blood alcohol levels of approximately 0.1 percent, a device that was different from the Multiple Task Performance Battery described above was used, but the requirements for time sharing were similar; performance of different combinations of mental arithmetic, monitoring, and two-dimensional tracking tasks was required (15). The results showed that the monitoring

tasks were affected at each of the two levels of workload used, but the tracking task was affected only at the higher of the two workloads (tracking, monitoring, and arithmetic). The arithmetic task was not significantly affected under either workload condition. In this study, the subjects apparently regarded the arithmetic task as being a "primary" task and gave it priority over the other tasks; it could perhaps be argued that the subjects "protected" their arithmetic performance at the expense of the other tasks. When just the tracking and monitoring tasks were presented, it could similarly be argued that they placed priority on the tracking task and "protected" that performance. Whether or not these proposed interpretations are accepted as reasonable, it seems clear (and very commonsensical) that the priority an operator assigns to a task will be an important factor in determining the level of performance maintained on that task as other duties are added.

IV. Analytic and Synthetic Methods.

The methods to be discussed in this section have been somewhat arbitrarily categorized as analytic or synthetic. (Both types of methods have some elements of each general approach, but, in my opinion, the first to be discussed leans a little more in the analytic direction and the second, a little more in the synthetic direction.)

A. Analytic Method. Senders has been a major proponent of the analytic method of workload analysis (44,45,46,47,48). This basic approach rests on the following assumptions (49):

1. Visual distribution of attention is the major indicator of operator workload.

2. The various signals that must be monitored demand attention commensurate with the characteristics of the signal and the required precision of readout of the signal by the human operator.

3. The human operator is effectively a single-channel device capable of attending to only one signal at any time.

4. The probability of human failure at any time is equal to the probability that two or more signals will demand simultaneous attention.

Senders states that these are simplistic assumptions in the sense that other signal sources (e.g., auditory) are not considered; attention to the visual part of continuous manual control tasks is not considered; and peripheral vision is not taken into account. Thus, the major analyses have to do primarily with instrument layout and deal only with requirements for instrument reading as a source of workload.

An important feature of this approach is that it can be applied in advance of the existence of specific hardware; it requires only that certain conditions be specifiable. For a given visual display, if the following information is available, then workload-related parameters can be calculated:

1. The maximum or cutoff frequency of the display must be specified. From this figure, the required fixation frequency as a function of time can be calculated.

2. Signal amplitude and acceptable error of reading must be specified.

From 1 and 2 the information rate for the display can be calculated. From the information rate, the fixation duration can be calculated (on the basis of the known relation between information content and response time). The product of fixation frequency and duration of observation yields the time required for observing the display expressed as seconds/second. The times found for each display instrument can be summed to get an index of monitoring workload as total seconds/second required overall in observing instruments. If uncorrelated signal sources are assumed, transition probabilities (e.g., probability of looking at display B after having observed display A) can be calculated and thus lead to guidelines for optimum instrument layout.

Senders (46) tested these notions in a laboratory situation by using four meters that were driven at different frequencies. He then compared predicted fixation frequencies based on the display characteristics with fixation frequencies as determined by motion pictures of the eye positions of the subjects. The agreement between prediction and data was quite good. Subsequently, Carbonell, Ward, and Senders (9) compared predictions with data from pilots flying approaches to landing in a simulator. Instrument pickoffs were used to establish the frequency characteristics of the various instrument displays and eye-movement measures were used to determine fixation frequencies. The agreement between the values from the prediction procedures (Nyquist model) previously used (49) and the data was reasonably good; however, a queueing theory model gave substantially better agreement.

Clement, Jex, and Graham (17) describe the application of a "manual control-display theory" to instrument landings of a "large subsonic jet transport." This theory, detailed by McRuer and Jex (37) and McRuer, Jex, Clement, and Graham (38), attempts to use hypothesized ratios between fixation frequencies and display bandwidths that are tailored to the accuracy-of-control requirements for the particular display. Then, using a procedure otherwise similar to that described by Senders (49), Clement et al. (17) computed a fractional scanning workload index for each display function and summed these arithmetically to get a quantity that is equivalent to a seconds/second scanning index. They showed that, as a design exercise, the predicted scanning workload for a selected aircraft panel layout could be reduced from 1.32 (anything greater than 1.0 is overload) to 1.01 by

combining certain displays. Although their predicted best display arrangement "agrees with that actually adopted" by a major airline for FAA Category II certification, empirical validations of scan times and fixation durations are not presented. In a subsequent study, Weir and Klein (54) collected data by using a "DC-8" flight simulator; however, their results in terms of scan times were compared with previous findings with aircraft and simulators rather than with theoretical predictions based on display information. Further discussion of this analytic approach can be found in Allen, Clement, and Jex (1).

The analytic approach to workload prediction requires considerable knowledge about the characteristics of the forcing functions of the various instruments and displays. But, where such information is available, the methodology developed to date shows promise, especially in applications to new, design-stage systems. However, substantial effort in the empirical validation of the procedures is still needed and warranted.

B. Synthetic Method. What is being referred to here as the synthetic method might equally well be called a combinatorial method. The point of departure of this method is a task analysis of the system; the proposed mission or operating profile is broken down into segments or phases that are relatively homogeneous with respect to the way the system is expected to operate. For each such mission phase, the specific performance demands placed on the operator are identified through task analysis procedures. Once individual tasks and subtasks have been isolated, previously available (e.g., Munger, Smith, and Payne, 40) or ad hoc data are compiled on the performance of the tasks with both performance times and operator reliabilities being taken into account. The information on performance times is then accumulated for a given mission phase and the resultant sum is compared with the predicted duration of the phase. The comparison of these two quantities--time required to perform versus time available--can be used to reflect an index of workload. Although other factors can be included in this synthesizing process, time is typically the primary variable considered.

One example of this approach is the Cockpit Evaluation and Design Analysis System described by Brown, Stone, and Pearce (8). Brown et al. define workload as follows: "Flight crew workload is the ratio of the summation of required crew-equipment performance time to the time available within the constraints regulated by a given flight or mission." Their design and analysis system is computerized and is organized in such a way that detailed information can be included regarding required times, available times, items of equipment involved, and flight phases as well as the design personnel responsible for the various equipments and subsystems.

Flight phases are further broken down by identification of what they call milestones, a milestone being a change in heading, airspeed, altitude, etc. Preliminary allocations of duties and activities are based on operating techniques of expert pilots and operating procedures for similar aircraft. For purposes of workload prediction for a given segment, the computer output

is expressed in the form of percentage-of-capacity figures for each task element each crew member is to perform. In this way critical periods in a mission phase can be identified and possible corrective measures evaluated. The primary purpose of the design analysis system " . . . is to provide data for use in comparative evaluation of alternative crew station designs." Its major values are the ease with which system changes can be evaluated. As Brown et al. state: "Any workload reduction must be evaluated in terms of the context within which this occurs and it seems senseless to increase cost by automating a feature that saves work during low workload periods only."

There are a number of other instances of the application of the synthetic methodology to the problems of workload prediction. Although the basic approaches are similar, there are some potentially important differences in detail. For example, Klein and Cassidy (32) describe an approach to estimating work requirements in which, apparently, an average required performance time is used to reflect the contribution of each task to the total work requirements, but the sum of these times can exceed the time available and thus lead to the notion of time stress. Their general procedure for analyzing the mission requirements is basically as described above. Klein and Cassidy also point out the need to recognize the nonadditivity of workload elements. This nonadditivity was investigated by evaluating a tracking task when performed in conjunction with a discrete task; they concluded: "Workload elements do not interlace in a directly additive fashion."

Wingert (56) places considerable emphasis on the fact that the performance of two tasks in combination often represents a workload that is less than the sum of the individual workloads. He used a model that took account of the nature of the task input (visual, auditory, or kinesthetic) and the task output (motor, vocal, or none required). He then prepared an "interlace table" for different combinations of two tasks with the various possible combinations of input and output modes. The actual values used in the table depended on analyses of the scanning requirements, information-processing-time predictions, and the set of summation rules assumed to apply to particular pairs of inputs and outputs. A specific set of tasks was evaluated by using a fixed-base helicopter simulator, and "interlace coefficients" were determined. The resultant coefficients are used, in the simple case, as follows:

$$\text{Total workload} = WL(1) + WL(2) - I * WL(2)$$

where I = the interlace coefficient.

Wingert discusses the concept of interlacing in the context of parallel versus serial processing of information, and, in general, the amount of interlacing expected depends on the extent to which parallel processing is possible.

This notion of interlacing can also be viewed from the simpler time-sharing frame of reference. The highly skilled operator has typically "automated" many aspects of his complex task in a way such that many of the elements require little if any information processing (channel capacity) for satisfactory execution of the required behaviors. Consider a two-dimensional tracking task as represented by the instrument landing system (ILS) display. Assume that the pilot, on approaching the outer marker, observes that he is slightly (but undesirably) below glide slope. Through long experience, he is able to apply an appropriate adjustment that will bring the aircraft smoothly to the glide slope. He does not then sit and watch the needle slowly drop! He turns his attention to other displays (e.g., airspeed) and knows approximately when to return his attention to the ILS display. Similarly, once he has the ILS needles centered and has established a proper rate of descent, only under very adverse conditions of wind and buffeting will he have to give the ILS display his undivided attention. In other words, how often he must look at a display to insure satisfactory performance depends on the "forcing function" acting on that display and the criticality of the task in terms of permissible error rates and amplitudes (cf. Senders, 49). To consider another kind of behavior, the neophyte automobile driver must give most of his attention to the steering task of "keeping the car on the road." For the expert driver, steering is concerned with avoiding rough spots, maintaining safe separations from oncoming traffic, etc.; keeping the car on the road has been automated. And if we look far enough we may run across an oldtime telegraph operator who can send or receive a message while simultaneously telling us about the good old days.

However, we should keep in mind that, at least at the present state-of-the-art, caution is in order in assuming too much interlacing. Such skills may be highly vulnerable to stress and other such factors (13). By way of analogy, we do not want an aircraft designed to just withstand the maximum expected g and gust loads.

V. Simulation Methods.

A. Fidelity. Webster (53) defines a simulator as "one that simulates, specif: a device in a laboratory that enables the operator to reproduce under test conditions phenomena likely to occur in actual performance." If we interpret the word "phenomena" to mean "system-operating characteristics," then the dictionary definition certainly states the intent of the designer of the simulator. Chapanis (10) considers a simulation to be a kind of model and prefers to define models as simply being analogies of some particular part of the real world that is of interest to the model maker. Chapanis makes a good case for this usage, and an important value in thinking of a simulation as being an analogy is that we are all aware that analogies tend to come apart when they are pushed too hard or are examined too closely. When we talk about fidelity of simulation, we are thus talking about "how hard we can push" before the analogy breaks down.

The difficulties encountered in achieving adequate fidelity in a simulator are primarily a function of the purpose for which the simulator is to be used. Thus, for some purposes, a control stick and a display with an appropriate interface provide adequate levels of fidelity. As Hopkins (28) has said, the kinds of things that are needed on a simulator depend on "(1) your purpose in using it, and (2) your method of using it. . . . Cost effectiveness has not been demonstrated for all the bells and whistles that come as standard trimmings on our current flight training simulators."

B. Assumptions. The basic assumption underlying the use of simulation in virtually any context is that the device represents to a satisfactory degree those elements of the system being simulated that are important and relevant to the purposes of the enterprise being undertaken. More specifically, in using a simulator to study pilot workload, it is assumed that:

1. Those factors in the real system that are relevant and important to the operator functions being evaluated are present.

2. Those aspects of the simulation that differ from the real system will not introduce important disturbances in the measures being taken.

3. Behavioral effects of task manipulations can be isolated from simulator operating characteristics as sources of variance.

4. The performance effects of the variables being manipulated in the simulation do not importantly differ from the effects that would occur in the real system.

Most of he work that has focused on the evaluation of the usefulness of simulators has been done in the context of the substitution of simulator training or experience for actual flight training or experience, and even in this area many questions regarding training simulators have been at best only partially answered. (A special issue of Human Factors (1963, No. 6) was devoted to this problem area.)

Unfortunately, many of the investigations that have looked at workload and other design questions using simulation have been reported in private company or laboratory internal publications or not at all. Thus, the open literature is virtually devoid of well-documented studies in which simulation--in the ordinary meaning of that term--was used to investigate workload; i.e., where measures were taken from the simulator to provide indices of the performance effects of workload variations as produced by changes in the simulator tasks.

C. A Flight Simulator Example. Corkindale (20) reported a study of missile control performance as a function of concurrent workload using a fixed-base flight simulator. The study included the following workload conditions:

1. Missile control tasks only. (Two-dimensional tracking using a joy stick with the left hand and a TV display.)

2. Simulator manual control using a Head-Up Display (HUD). (Two-dimensional tracking with control column.)

3. Missile control plus HUD manual control. (Two, independent, two-dimensional tracking tasks--one with left hand and one with right hand. At the end of first 90 seconds, the TV came on and the subject watched for appearance of target.)

4. Missile control task plus HUD monitoring. (Two-dimensional tracking of missile plus monitoring of HUD for an infrequently presented signal that subject responded to by pressing a button on the control column.)

Performance of the missile and aircraft control tasks was measured by recording integrated error in each axis for each tracking task. In addition, detection time for the TV target was measured. Once the TV target was acknowledged and the crosshairs had appeared, the missile tracking task lasted just 10 seconds; the HUD aircraft control task, when present, lasted for approximately 3 minutes 10 seconds; the missile control task always fell in the second half of the test trial.

All but one of the measures evaluated were significantly affected by workload; surprisingly, horizontal error in tracking the TV display target was not sensitive to these workload variations. A major conclusion drawn by Corkindale (20) was that his findings fit well with the work that Rolfe (42) reviewed and interpreted to indicate that secondary tasks typically produce degradation of the performance of the primary task in spite of instructions to maintain the highest level of performance on that task. It would be interesting to know what sort of prediction the analytic method of estimating workload (e.g., Senders, 49) would make as regards the task combinations used. Corkindale cites evidence that the subjects spent a significantly smaller percentage of the time looking at the HUD when the TV was on (29.3 percent) than when the TV was off (60.3 percent), even though the HUD was the primary source of feedback to the subject as to how well he was controlling the aircraft. Therefore, one would be tempted to speculate that the analytic method would predict that a pilot cannot do both of the tasks without at least some degradation of performance on both. What, then, should we expect the pilot to do when we ask him to try to do both tasks simultaneously? Assuming that the pilots used in such a study were mission oriented, then their approach to the situation might very well be as follows:

> "This is an exercise in which I am expected to hit a target with an air-to-surface guided weapon. I have to control the missile and fly the airplane. I know that I cannot fly as well while controlling the missile as I can while I am not. So, I will try my best to hit the target and will consider the mission a success if I score a hit and do not crash."

It could be argued that many military pilots would follow this line of reasoning unless they were told that they must maintain undiminished control of the aircraft even if they never hit any targets. And with instructions of that sort, it might be difficult to maintain good levels of subject motivation to perform the task.

Assuming that Corkindale's subjects were able to handle the aircraft control task in a manner that satisfied them when that was their only task, what does a (significant) doubling of the error scores with the addition of the TV task mean? Did the pilots think they were controlling the aircraft in an acceptable manner in the two-task condition? Whether they did or not, what was their criterion? Did any of them ever "crash"? Without some sort of absolute error criterion, the interpretation of the results in this kind of study (or any simulator study) is very difficult. We are on somewhat firmer ground if the purpose of a study is to compare the workload properties of, for example, two alternative ways of displaying the same information. If there is a substantial and statistically significant advantage of one alternative, then cost-versus-effectiveness analyses can be made. But even in this simpler case, the absence of absolute criteria creates problems; for example, what procedure can be used to establish what a "substantial advantage" is in relation to "real world" requirements? In other words, we must not forget that in many important respects a simulation is merely an analog of some aspect of the real world.

D. A Space Simulator Example. Cotterman and Wood (21) attempted a direct treatment of the problem of criteria in a simulation context in a study of the retention of pilot skills associated with a lunar landing mission. This study involved a full mission simulation at the Martin-Marietta Corporation as part of the NASA space program. The subjects in this study were 12 aerospace research pilots who had participated previously in a Human Reliability Program study conducted with this simulation system. The specific goal of the study reported by Cotterman and Wood was the evaluation of the retention of skill after relatively long periods (13 weeks) of disuse. The total study concerned nine separate mission phases, with from one to four performance criteria for each phase. For present purposes, only one phase will be discussed; viz, the "Brake and Hover" phase involved in the lunar landing.

Based on engineering analyses, permissible error rates had been established for four motion parameters during the Brake and Hover phase. These were: displacement (or range error), 200 feet; displacement rate, 10 feet/second; impact rate, 10 feet/second; percentage fuel consumed, 95 percent. Exceeding these values by appreciable amounts would incur unacceptable risk of mission failure.

The analytical approach applied by Cotterman and Wood was to use the data on the last four training trials for each pilot to establish a mean and a standard deviation for each parameter. Since their interest was in

establishing whether subjects could attain performance at a high level of consistency, they selected a statistical criterion that was associated with a probability of 0.950 that the subject would perform within the criterion tolerances. The actual calculations, though somewhat laborious if done by hand, are conceptually simple. First, the standard deviation for the data from a given pilot for a given measure is computed; then, a normal deviate ("z" score) is found by dividing the difference between the criterion and the obtained score of interest by the standard deviation. A table of normal deviates can then be used to establish an approximation of the probability that the pilot in question will in fact be expected to stay within the criterion, or, using the appropriate equations, an exact probability can be computed. For one subject in the study reported by Cotterman and Wood, it was found that probabilities of staying within the criterion on the four previously mentioned variables were: 0.998; 0.525; 0.9995; and 0.9995. If the events on which each of these probabilities is based are independent, then their cumulative product is the probability that the entire mission phase will be within the criterion limit. With this approach, whether applied to simulation or to an in-flight situation--assuming that the criteria can be specified--the probabilities can be developed in a way that makes them useable for purposes of reliability engineering. The requirements are (1) the data must be quantitative in form, (2) enough repetitions per subject must be provided to achieve reasonably reliable estimates of the standard deviation, and (3) some criterion must be available that is specifiable in quantitative form.

VI. In-Flight Methods.

A. System-Based Measures. Various techniques have been used to record indices of performance in aircraft. They have involved varying degrees of difficulty of installation and have been used with varying degrees of success. Some of the earliest systems used voltage analogs, either from direct instrument pickoffs or from repeater instruments, to drive the pens of an ink-writing oscillograph. More recently, frequency modulation techniques have been used to record analog signals onto magnetic tape; off-line computer readout and analysis can then be applied to the tapes. And still more recently, on-board digitizing techniques have been used to record data on magnetic tape directly in digital format for later computer analysis.

Some of the earliest work on studies of aircrew workload involved variations on the standard techniques of time-and-motion study (e.g., Christensen, 16), and, at about that same time, pilot workload (instrument scanning) was studied by use of motion pictures of pilot eye-movements during instrument approaches (39). Still more recently, Weir and Klein (54) describe the use of an Eye-Point of Regard system that uses a horizontal movement detector (bite board) and corneal reflection to give a resolution of "about $\pm$ 1°" in either axis with respect to the eye fixation point. Photographic and videotape techniques have also been used to record general pilot

activities in simulators as well as aircraft; e.g., time/frequency measures of control usage. And still more recently, Geiselhart, Shiffler, and Ivey (24) used time-and-motion study techniques in evaluating crew requirements for the KC 135 tanker aircraft on actual missions.

Roscoe and Williges (43) reported a study carried out in a Beechcraft C45H using each of eight experimental display conditions under simulated instrument flight conditions. The tasks confronting the subjects, who were naive to flying, were (1) tracking a randomly generated command flight path; (2) a disturbed attitude task that required subjects to compensate for Gaussian noise summed with the actual bank attitude signal; and (3) recovery from unusual attitudes entered with subliminal angular accelerations. All data were recorded on a strip recorder and on magnetic tape. Among other results reported by Roscoe and Williges was the finding that the maintenance of command heading was significantly better with the displays in a pursuit mode as compared to a compensatory mode.

Knoop (33) reports a study designed to evaluate the feasibility of automatically assessing T-37 student pilot performance in the Air Force Undergraduate Pilot Training program. A T-37B aircraft was instrumented to record 21 flight and control parameters in digital form on magnetic tape. Major variables (airspeed, pitch, roll, stick position in two dimensions, and rudder position) were sampled 100 times per second. Other variables, such as altitude, heading, flap position, etc., were sampled at a 10-Hz rate. A major part of this effort involved attempts on the part of instructor pilots to fly prescribed maneuvers in as nearly perfect a manner as possible. These maneuvers were broken down into phases and subjected to computer analyses in an attempt to develop measures that best characterized a high level of performance; concurrently, subjective ratings of the instructor pilots were also used as part of the evaluation. The resultant functions of the various control and performance parameters were compared with those of student pilots to try to identify those measures that best discriminate between trainees and skilled pilots. Overall, this effort met with mixed success, and major attention was diverted to trying to follow the progress of students through the training program. A major difficulty encountered was the clear lack of agreement across instructors as to what was most important in characterizing good performance in particular maneuvers.

Hasbrook, Rasmussen, and Willis (27) reported an in-flight evaluation of a "peripheral vision flight display" (PVFD) in a Beechcraft Bonanza 35A aircraft. Each of 20 pilots flew two ILS approaches with a conventional display system; they also flew five approaches with the PVFD system, but only the last two of these approaches were considered for data analysis purposes. Performance levels were recorded on a 14-channel FM analog tape system installed in the left rear seat of the aircraft. Twelve channels of information were recorded: pilot heart rate; aircraft pitch and roll (taken from the primary attitude indicator); vertical and lateral deviations from the ILS centerline (taken from the glide slope and localizer signals);

altitude, airspeed, and vertical speed (obtained from the aircraft's pressure and static air systems); vertical acceleration (taken from an accelerometer located near the center of gravity of the aircraft); heading deviations (taken from a remote gyro-stabilized compass); and control wheel data (derived from mechanoelectric transducers connected to the aircraft's control cables). Event signals were inserted on a separate data channel by the use of a manual switch. Data were recorded starting at the beginning of the approach at the outer marker and ending when the runway threshold was crossed at an altitude of 100 feet; at that point the subject was instructed to increase power and go around. No differences were found between the displays, but the more experienced pilots of the group (an average of 1,267 hours of instrument time) maintained a small, significant superiority on holding to the glide slope between the outer and middle markers as compared to a less experienced group (an average of 104 hours of instrument time). Thus, although Hasbrook et al. stated that the pilots generally rated the PVFD as good to excellent, the PVFD display configuration did not result in statistically superior performance.

Billings, Gerke, and Wick (6) did a study that, though it did not involve manipulation of workload, is of interest because it involved both in-flight and simulator performance. The variable of interest was the dosage of sodium secobarbital (0, 100, or 200 mg). The in-flight portion of the study was carried out by using a specially instrumented Cessna 172; the simulator part of the study used a GAT-1 simulator. For both the aircraft and the simulator, data were recorded in digital format at a sampling rate of 25 Hz to yield measures of average absolute error in holding to the localizer, glide path, and commanded airspeed (100 mph); root-mean-square (RMS) error was derived by appropriate computational procedures for each of the variables. The five "highly experienced professional pilots" who served in the study showed a small, nonsignificant overall increase in error across the six aircraft flights (averaged over drug conditions) and a slightly larger, significant decrease in error over the six simulator flights (again averaged over drug effects). It is interesting to note that whereas all of the six statistical tests carried out on the simulator data showed a significant drug effect, only four of the six tests on the aircraft data showed the drug effect to be significant. In addition, for all segments of the approach the no-drug (placebo) condition was best in the simulator, and for all but one segment the 100-mg dose resulted in better performance than did the 200-mg dose in the simulator. The analogous results were mixed in the case of the aircraft data. On all three measures (glide slope, localizer, and airspeed) the RMS variability was less in the simulator than in the aircraft; and for only one absolute measure (deviation from command airspeed at the 200-mg dose) was performance in the simulator numerically poorer than in the aircraft. Direct statistical comparisons between simulator and aircraft were not reported; perhaps they were not feasible.

B. Externally Based Measures. Brictson, Ciavarelli, and Wulfeck (7) describe a system that has been used to assess the quality of aircraft carrier approaches and landings. The workload variations were those associated with night versus day landings. The procedure for recording the final approach performance involved a shipboard instrumentation system consisting of twin precision radars and a signal data recorder that provided up to eight channels of continuous flight information. The range error was reported to be on the order of 4 feet and the angular error, on the order of 0.3 milliradian. Range, true altitude, altitude error, lateral error, sink speed, true air speed, deck pitch, and closing speed were the variables usually recorded. Among other findings, Brictson et al. reported that altitude errors were greater at night than during the day with a greater tendency for the approach to be below glide slope at night. Brictson et al. also report that a reasonably good measure of the quality of the approach and landing was obtained by simply noting which of the four arresting wires was hooked and the number of "bolters" (no arresting wire engaged). The major difference in the tasks of night versus day landing was in the impoverishment of the visual field in terms of details of the carrier and the texture of the water. Not having those cues made the task more difficult, and Brictson et al. were able to develop differential criteria for predicting successful landings at night versus during the day for various departures from the optimum approach configuration.

VII. Discussion, Recommendations, Cautions, and Conclusions.

A. A Hypothetical Research Vehicle. Let us assume that there exists a real aircraft system with the following capabilities: (1) An exact assignment of the nature and number of pilot duties or activities can be made for any given mission phase. (2) It is possible to vary those duties singly or in combination over time. (3) Control and display characteristics can be manipulated at will. (4) Precise and reliable quantitative indices of the task demands placed on the pilot by the system are available for all task elements. (5) Precise and reliable quantitative measures of the skill with which the pilot meets those demands are available. (6) An adequate criterion measure of system performance is available.

What kinds of information might we expect to be able to develop as regards pilot workload through use of such a system? First, as we add tasks in different combinations, we should be able to determine the priorities the pilot assigns to the different tasks and whether these priority assignments are consistent across pilots; as the number of actions required per unit of time approaches and exceeds the time available, or as simultaneous demands for action arise, some tasks will be given less attention with a resultant lowering of performance on those tasks. Second, we should be able to determine how the different elements of the pilot's job interact; as different tasks are added to the total workload, do some tasks tend to interfere with the performance of other tasks? And, third, we should be able to determine what kinds of tasks or performance functions are most sensitive to variations in total demand.

In a similar manner, for a given task load on our assumed system, we should be able to determine the relative sensitivity of the different performance demands to various environmental and procedural factors. We should, in this somewhat different context, again see which tasks are given priority. And we should be able to acquire information on the relative importance of "operator style" in system performance.

From systematic studies of task characteristics, task combinations, and procedural factors, we should be able to develop a quantitative concept of workload capacity or--as some prefer to call it--channel capacity. Thus, we should be able to arrive at a notion of workload for a given mission phase as involving some portion of the pilot's total moment-to-moment capacity to satisfy the system demands.

Unfortunately, there appear to be no instances in which a system or a simulated system has been subjected to these sorts of manipulations in any kind of programatic attack on the nature of pilot workload. (Although something like this has been done with synthetic work tasks, the programs have not been as complete or as systematic as would be desirable, and the results are, therefore, of more relevance to environmental and procedural variables than to workload per se (cf. Chiles et al., 13; Alluisi, 2).)

However, we can, perhaps, make some empirically based projections (educated guesses) as to what some of the products of such a program might be. First, we would surely find that some tasks will be given priority. Which ones will depend on training and the perceived criticality of the task to the safety of the system and to the probability of mission accomplishment. For example, ILS-type guidance information will be given very high priority during very low visibility approach conditions; and there is reason to believe that some of the instruments are, on occasion, given too low a priority after breakout with potentially disastrous results.

Another predictable result is that the elements of many combinations of tasks will be found to be nonadditive (in the simplest meaning of that term). At high levels of pilot skill at time sharing, a number of tasks can apparently be performed without evidence of decrements or cross interference. However, where tasks present conflicting demands, the lack of additivity may take on a much different character; the specific effects will largely depend on the required sampling rate for the different information sources coupled with the required "dwell times"; i.e., how long it takes the pilot to extract the necessary information. Perhaps the most important single factor in this area is the degree of freedom the pilot can exercise as to exactly when various actions must be initiated.

If the suggested program were to be carried far enough, it would probably develop that only a limited number of operator styles will emerge that will allow or insure overall satisfaction of the system demands.

And, finally, it will be only after substantial and thorough research that the quantitative methods will yield readily useable indices that relate directly to "how hard the pilot has to work" with a given system workload configuration.

The fact that these above-mentioned "educated guesses" are, for the most part, rather obvious should not be allowed to detract from the clear desirability of attempting their empirical verification. Perhaps on such a "bare bones" kind of outline a general theory of workload could be developed.

B. Choosing a Method. The first and foremost factor to keep in mind in choosing a methodology in attacking some particular workload question is the purpose or goal of the research. This is true whether we are choosing from among the kinds of methods discussed here or from among those discussed in some other study.

The primary thing to keep in mind is that the measures being taken should allow the detection of operationally important changes in the pilot's ability to satisfy system demands as a function of the workload variables being manipulated. If a given measure or pattern of measures were to reveal decrements for one configuration of system demands in relation to another configuration, the decrements should be meaningfully relatable to critical operational tasks in terms of pilot reliability, system safety, and/or probability of mission success. Alternatively, (and this is much more difficult to establish) if no decrements are found for a given workload configuration, it should be clearly possible to predict that the pilot could satisfy the system demands under operational conditions. At the same time, every possible effort (within reason and the scope of available resources) should be made to design the research so that maximum generality across systems is possible. Clearly, when we choose a method and select the variables that are to be measured (the dependent variables), we are committing ourselves to a particular realm of discourse as regards system workload parameters. Thus, we must be certain that the basic problem that gave rise to the research can in fact be handled within that realm of discourse. (The importance of the selection of dependent variables has been dealt with in some detail by Chapanis, 11; by Alluisi, 3; and by Chiles, 12,14.)

The most pressing and the most difficult problem in assessing workload effects (whatever method is chosen) lies in the development of reliable, quantitative criteria that validly reflect system performance. We need criteria against which to evaluate the results of our research. We must be able to distinguish acceptable from unacceptable, good from acceptable, and excellent from good performance of the system. We must be able to make these distinctions quantitatively and reliably. And we must be able to disentangle pilot performance, machine performance, and pilot-machine performance. Ultimately, we want a method with which it would be possible to assign reliable variance, as appropriate, to the man, to the machine, and/or to the man-machine interface.

For some specific questions this may appear to be a deceptively approachable question. For example, if we need to determine which of two instrument landing systems makes the smaller contribution to pilot workload, we could simply secure accurate measures of the deviation of the aircraft from the glide slope and the localizer and perhaps monitor airspeed. Comparison of the values of these measures for the two displays should give us an index of their workload-inducing properties. However, it is entirely conceivable that one display would lead to smaller errors only because the pilot could, by working harder, take advantage of some peculiarity of that display in holding to the proper course; at the same time, the pilot might very well be less able to respond appropriately to some emergency condition that might arise from some other quarter. Thus, in this specific example, we would need to add a variable that would shed light on how much of the pilot's workload capacity was being used up by each display. In our hypothetical, completely flexible aircraft system, we could introduce some sort of malfunction that, conceivably, could be handled readily with the otherwise poorer display but only with considerable difficulty in the case of the "better" display. This is admittedly a highly artificial example and the intent is merely to suggest a possible way in which what might appear to be a simple measurement problem might not be so easy after all. The other intent in introducing the example is to suggest that when we draw a conclusion based on a particular set of measures, the results may imply extrapolations well beyond the circumstances under which the measurements were made. (Remember, analogies, as well as examples, should not be pushed too far.)

The measurement and analysis approach described by Cotterman and Wood (21) in their evaluation of performance in a space vehicle simulator appears to show considerable promise as a technique for converting "raw" performance measurements to probabilities of meeting criterion requirements. However, there is a gap between their application and the typical pilot workload measurement situation. Specifically, in the case of the Lunar Excursion Module, the maximum values of various parameters can be specified quite readily; for example, engineering specifications dictate that the impact velocity of the vehicle on landing cannot exceed some value without risk of damage. Such precision is less clearly identifiable in the majority of aircraft operating situations; typically, rather broad latitude is possible in the flight parameters without risk of entering unsafe conditions of flight. Thus, in some areas the application of the procedure to some aircraft mission phases might become a bit arbitrary. Perhaps for research purposes it would be necessary and profitable to set up much more stringent criteria than normal, but not too stringent; the difficulty of the criteria should be such that the typical pilot from the population of pilots to which we wish to generalize would, under normal conditions, be capable of performing satisfactorily.

Assuming that we have adequate criteria of system performance that reflect both man and man-machine contributions to system output, how do we proceed?

The first step is the identification of all of those human and machine factors that could conceivably influence the variable of interest. This list typically will be unmanageable from a research point of view, and expert judgment, based on knowledge of human behavior and system behavior, will have to be applied to eliminate those factors of negligible or relatively small potential impact. Having developed a (presumably manageable) list of important factors, we attempt to phrase (or rephrase) the question such that it becomes amenable to some (as yet unspecified) research technique. We next arrange the relevant factors into two categories; one category contains items that are in the nature of constraints or boundary conditions, and the second category contains items that are in the nature of possible independent variables; this second category will, of course, include the factor or factors that gave rise to the need for the research in the first place. Now we are ready to examine the situation in detail in order to make a decision as to what would be the best research methodology to apply to the problem. At this point the available guidelines become very ambiguous and professional judgment must play a dominant role.

First, we look at what are referred to above as the boundary conditions; these are the fixed aspects of the operational system from which the problem derives; they concern factors such as the gross weight of the vehicle, its flight range, mission characteristics, number of engines, etc. Each of these factors is evaluated in relation to the question: "Might this factor be reasonably expected to have an effect on the performance in question?" Then we examine each item on the list of possible independent variables; and again we ask the question: "Might this factor be reasonably expected to have an effect on the performance in question?" Depending on the pattern of "yeses" and "noes," we will tend to direct our attention toward one methodology or another.

If, for example, the basic problem is concerned with a perceptual question, say a visual discrimination in reading two different types of dial, and kinesthetic or gravitational cues would not be expected to play a role, then perhaps a more or less traditional laboratory study might be appropriate. (We will refer to this study as task A.) However, if the instrument reading must be made while performing some other task, say a two-dimensional tracking task (we will call this study task B), then perhaps a part-task simulator would be in order. If the performance of task B may be importantly influenced by the insertion of command information, then a more elaborate simulation might be in order (task C). And if kinesthetic cues may be important, we may need to go to a motion-type simulator or perhaps an in-flight evaluation.

Finally, we must select the dependent variable--the thing we are going to measure. This may be a time measure: how fast can the pilot do a task? It may be an absolute error measure: how often did he hit the wrong switch? It may be a relative error measure: what was his average deviation from glide slope? Whatever the measure, we should if at all possible try to relate the findings back to system-relevant criteria developed in a manner analogous to

that described by Cotterman and Wood (21). All too often, the thing that is chosen for measurement is that which is easiest to acquire or has been used most often in the past without any specific rationale having been shown that relates the measure to real-system performance questions.

In some cases the results of the study (accuracy of dial reading in the above-described example) may provide information that is more or less directly interpretable in terms of workload. But what if there is no change in any of the measures as a function of which dial is used? Can we infer that the two dials represent equal workload contributions? The answer is, of course, no. Only after we have pushed the total workload to a maximum reasonable and likely level and found no differences on any measures should we be willing to assume the equality of the two displays. (It is a peculiarity of statistical methodology that we cannot prove they are equal.) The procedure we use to push the apparent level of workload to a maximum is, again, a matter of professional judgment. But it is an extremely important judgment. If workload is added in an obviously artificial manner, especially if our subjects are operational personnel, we may lose them--motivationally speaking. We must always be sure that the research situation--be it laboratory, simulator, or aircraft--is presented in a manner such that it will be responded to as a "real" situation as opposed to a game or a contrived--and thus (perhaps) meaningless--exercise.

Let no one make the mistake of assuming that this process of choosing a method is easily executed. The problems are many and the decisions difficult.

C. Conclusions. The general approaches that we have labeled "laboratory methods" are probably best suited to conducting background research on more general questions pertaining to workload. Wherever they are appropriate they are the method of choice because of the typically high degree of control possible and the attendant high levels of reliability. The synthetic work method is especially well suited to examining general workload questions because, by its nature, tasks can be added, removed, and modified with relative ease, and, depending on the overall level of complexity, large investments in training time are not required. The fact that it does not simulate an aircraft is both a strength and a weakness; it is a weakness because of problems of generalizing to specific systems; it is a strength because, if the tasks are well chosen, operational subjects can fairly easily be convinced to react to the synthetic work device for what it is and not make unfavorable comparisons between its behavior and the behavior of an aircraft. The secondary loading task method, especially when applied in a simulation or in-flight context, must be used with care. First, the task that is used to produce the load increments must be somehow (at least rationally) relatable to the kinds of activities it is presumed to assess in relation to the real system. Second, the properties of this task itself must be examined; at a minimum its reliability and relation to other tasks should be known. Although some authors (e.g., Rolfe, 42, and Corkindale, 20) argue that the primary task should remain unaffected by the introduction of the loading task, this

condition appears to be unnecessarily restrictive. If the loading task is properly selected (as noted above) and contradictory results are obtained (e.g., primary task A shows a decrement, primary task B is unchanged, but the loading task shows a decrement with B and not with A), the findings may be of little relevance to workload (or channel) capacity as a unitary concept; however, if such results were not simply the product of some uncontrolled condition, the finding would certainly be of theoretical if not practical interest. Perhaps it is better at this stage of development to consider the concepts channel capacity and single channeledness as being merely manners of speaking and serving primarily as heuristic devices. Although this does not argue against the ultimate possibility that the operator is single channeled, present evidence suggests that the information-handling capacity of the human operator is influenced by too great a variety of factors to try to permanently settle the single-channel hypothesis at this time. Returning to and slightly changing the above example, if task A shows more decrement than task B with the addition of the loading task and the loading task is performed better with task B than with task A, we certainly have learned something about the workload properties of the tasks. The findings, of course, remain ambiguous as regards channel capacity.

The analytic and the synthetic methods both appear to yield reasonable results, but both techniques rest on relatively fragile data bases. With further research on what I would call time sharing behavior, or what Wingert (56) calls function interlacing, the synthetic method promises to be a very useful aid in the design of systems and the allocation of workload. There is, however, considerable risk that the detailed task information required to apply the method will be collected and stored in a manner that will tend to limit its distribution and result in substantial amounts of unnecessary duplication of effort. Previous attempts to develop clearing houses for the information have not met with noteworthy success.

Simulators, especially those controlled by general purpose digital computers, have the potential of generating large amounts of very useful information on workload. However, whether the programs that resulted in their acquisition will allow adequate access to such systems for research purposes remains to be seen. But even given adequate access, research with simulators is not without its problems. First, naive subjects cannot be expected to learn to fly in a matter of a few hours; therefore, for most purposes--or at least for those purposes in which the full capability of the simulator is used--trained pilots are required who have adequate experience with that simulator and/or the aircraft it simulates. Thus, salaries can become a significant part of any substantial research effort. Second, the simulator is, first and foremost, designed and built to appear to behave like the aircraft it simulates; the quality of the signals internal to the simulator need not be very high to satisfy that requirement. Thus, especially with the older simulators, the available signals often introduce an unacceptably high degree of unreliability in the final measures. Third, because the simulator is designed to mimic the airplane, many of the functions are interconnected

in such a way that it can be very difficult to separate them out. For example, the relative contributions of the simulator, present performance of interest, concurrent performance that is not of direct interest and the interactions of these factors as sources of variance may be hopelessly entangled. And, fourth, also because the simulator is designed to mimic a particular airplane, generalization to other aircraft with significantly different characteristics (such as panel layout and operating procedures) becomes rather difficult.

Except for some of the safety limitations, in-flight methods can be used on virtually any problem suitable for investigation in a simulator. However, the recording of data of demonstrated reliability is a significant problem. Generally speaking, aircraft are electrically very noisy, and, where magnetic tape recordings are made (either digitally or through frequency modulation techniques), substantial programing for signal "reconditioning" is typically required; glitches are a constant source of annoyance (Knoop, 33). Unfortunately, no reports of reliability data have been discovered for in-flight recorded performance measures or for simulator performance measures. In fact, this is a major technical deficiency in virtually all the reported research using these two methods. (*This criticism applies equally well to* much of the other reported research related to the measurement of workload; viz, laboratory research.)

Some readers may be disappointed that firmer guidelines have not been offered as to how to design and conduct research on workload problems in aviation operations. Those who are familiar with the behavioral literature on the measurement of complex human performance will understand the absence of precise, "cookbook" rules for proceeding.

References

1. Allen, R. W., W. F. Clement, and H. F. Jex: Research on Display Scanning, Sampling, and Reconstruction Using Separate Main and Secondary Tasks. National Aeronautics and Space Administration Report No. NASA-CR-1569, 1970.

2. Alluisi, E. A.: Methodology in the Use of Synthetic Tasks to Assess Complex Performance. HUMAN FACTORS, 9:375-384, 1967.

3. Alluisi, E. A.: Optimum Uses of Psychobiological, Sensorimotor, and Performance Measurement Strategies. HUMAN FACTORS, 17:309-320, 1975.

4. Bartlett, Sir Frederick: The Measurement of Human Skill. BRITISH MEDICAL JOURNAL, 4510:835-838 and 4511:877-880, 1947.

5. Benson, A. J., H. F. Huddleston, and J. M. Rolfe: A Psychophysiological Study of Compensatory Tracking on a Digital Display. HUMAN FACTORS, 7:457-472, 1965.

6. Billings, C. E., R. J. Gerke, and R. L. Wick: Comparisons of Pilot Performance During Simulated and Actual Flight. AVIATION, SPACE, AND ENVIRONMENTAL MEDICINE, 46:304-308, 1975.

7. Brictson, C. A., A. P. Ciavarelli, and J. W. Wulfeck: Operational Measures of Aircraft Carrier Landing System Performance. HUMAN FACTORS, 11:281-289, 1969.

8. Brow , E. L., G. Stone, and W. E. Pearce: Improving Cockpits Through Crew Workload Measurement. Douglas AC Corporation Report No. MDC 63-55, 1975.

9. Carbonell, J. R., J. L. Ward, and J. W. Senders: A Queueing Model of Visual Sampling Experimental Validation. IEEE Transactions on Man-Machine Systems, MMS-9, No. 3, pp. 82-87, 1968.

10. Chapanis, A.: Men, Machines, and Models. AMERICAN PSYCHOLOGIST, 16:113-131, 1961.

11. Chapanis, A.: The Search for Relevance in Applied Research. In W. T. Singleton, J. G. Fox, and D. Whitfield (Eds.): Measurement of Man at Work. London, Taylor and Francis Ltd., pp. 1-14, 1971.

12. Chiles, W. D.: Assessment of the Performance Effects of the Stresses of Space Flight. Wright-Patterson AFB, Ohio; Aerospace Medical Research Laboratories Report No. AMRL-TR-66-192, December 1966.

13. Chiles, W. D., E. A. Alluisi, and O. S. Adams: Work Schedules and Performance During Confinement. HUMAN FACTORS, 10:143-196, 1968.

14. Chiles, W. D.: Complex Performance: The Development of Research Criteria Applicable in the Real World. In W. T. Singleton, J. G. Fox, and D. Whitfield (Eds.): Measurement of Man At Work. London, Taylor and Francis, Ltd., pp. 159-164, 1971.

15. Chiles, W. D., and A. E. Jennings: Effects of Alcohol on Complex Performance. HUMAN FACTORS, 12:605-612, 1970.

16. Christensen, J. M.: Aerial Analysis of Navigator Duties With Special Reference to Equipment Design and Workspace Layout. II. Navigator and Radar Operator Duties During Three Arctic Missions. USAF Report No. MCREXD-694-15A, February 1948.

17. Clement, W. F., H. R. Jex, and D. Graham: A Manual Control-Display Theory Applied to Instrument Landings of a Jet Transport. IEEE Transactions on Man-Machine Systems, MMS-9, pp. 93-110, 1968.

18. Conrad, R.: Adaptation to Time in a Sensorimotor Skill. J. EXPERIMENTAL PSYCHOLOGY, 49:115-121, 1955.

19. Conrad, R.: The Timing of Signals in Skill. J. EXPERIMENTAL PSYCHOLOGY, 51:365-370, 1956.

20. Corkindale, K. G. G.: A Flight Simulator Study of Missile Control Performance as a Function of Concurrent Workload. In AGARD Conf. Proc. No. 146, Simulation and Study of High Workload Operations, A5-1 to A5-6, 1974.

21. Cotterman, T. E., and M. E. Wood: Retention of Simulated Lunar Landing Skills: A Test of Pilot Reliability. Wright-Patterson AFB, Ohio, Aerospace Medical Research Laboratories Report No. AMRL-TR-66-222, 1967.

22. Darley, C. F., R. L. Klatzky, and R. C. Atkinson: Effects of Memory Load on Reaction Time. J. EXPERIMENTAL PSYCHOLOGY, 96:232-234, 1972.

23. Gartner, W. B., and M. R. Murphy: Pilot Workload and Fatigue: A Critical Survey of Concepts and Assessment Techniques. National Aeronautics and Space Administration Report No. NASA-TN-D-8365, November 1976.

24. Geiselhart, R., R. J. Schiffler, and L. J. Ivey: A Study of Task Loading Using a Three Man Crew on a KC-135 Aircraft. Wright-Patterson AFB, Ohio, Aeronautical Systems Division Report No. ASD-TR-76-19, October 1976.

25. Gerathewohl, S. J.: Definition and Measurement of Perceptual and Mental Workload in Aircrews and Operators of Air Force Weapon Systems: A Status Report. AGARD Conf. Preprint No. 181, Higher Mental Functioning in Operational Environments. London, Harford House, 1975.

26. Hall, T. J., G. E. Passey, and T. W. Meighan: Performance of Vigilance and Monitoring Tasks as a Function of Workload. Wright-Patterson AFB, Ohio, Aerospace Medical Research Laboratories Report No. AMRL-TR-65-22, 1965.

27. Hasbrook, A. H., P. G. Rasmussen, and D. M. Willis: Pilot Performance and Heart Rate During In-flight Use of a Compact Instrument Display. FAA Office of Aviation Medicine Report No. FAA-AM-75-12, November 1975.

28. Hopkins, C. O.: How Much Should You Pay For That Box? HUMAN FACTORS, 17:533-541, 1975.

29. Jahns, D. W.: Operator Workload: What Is It And How Should It Be Measured? In K. D. Gross and J. J. McGrath (Eds.), Crew System Design. Santa Barbara, California, Anacapa Sciences, Inc., pp. 281-288, July 1973.

30. Jennings, A. E., and W. D. Chiles: An Investigation of Time-Sharing Ability as a Factor in Complex Performance. HUMAN FACTORS (In Press).

31. Kelley, C. R., and M. J. Wargo: Cross-adaptive Operator Loading Tasks. HUMAN FACTORS, 9:395-404, 1967.

32. Klein, T. J., and W. B. Cassidy: Relating Operator Capabilities to System Demands. Proceedings of the 16th Annual Meeting of the Human Factors Society, pp. 324-334, October 1972.

33. Knoop, P. A.: Advanced Instructional Provisions and Automated Performance Measurement. HUMAN FACTORS, 15:583-597, 1973.

34. Knowles, W. B., W. D. Garvey, and E. P. Newlin: The Effect of Speed and Load on Display-Control Relationships. J. EXPERIMENTAL PSYCHOLOGY, 46:65-75, 1953.

35. Knowles, W. B., and D. J. Rose: Manned Lunar Landing Simulation. Paper presented at IEEE National Winter Convention on Military Electronics, Los Angeles, California, 1963.

36. Knowles, W. B.: Operator Loading Tasks. HUMAN FACTORS, 5:155-161, 1963.

37. McRuer, D. T., and H. R. Jex: A Systems Analysis Theory of Manual Control Displays. Proc. 3rd Annual NASA-University Conf. on Manual Control, Report No. NASA-SP-144, pp. 9-28, 1967.

38. McRuer, D. T., H. R. Jex, W. F. Clement, and D. Graham: A Systems Analysis Theory for Display in Manual Control. Systems Technology, Inc., Technical Report No. TR-163-1, June 1968.

39. Milton, J. L., R. E. Jones, and P. M. Fitts: Eye Fixations of Aircraft Pilots: II. Frequency, Duration and Sequence of Fixations When Flying the USAF Instrument Low Approach System (ILAS). Wright-Patterson AFB, Ohio, Report No. USAF TR-5839, October 1949.

40. Munger, S. J., R. W. Smith, and D. Payne: An Index of Electronic Equipment Operability--Data Store. Pittsburg, Pennsylvania, American Institutes for Research, 1967.

41. O'Donnell, R. D.: Secondary Task Assessment of Cognitive Workload in Alternative Cockpit Configurations. AGARD Conf. Preprint No. 181, Higher Mental Functioning in Operational Environments. London, Harford House, 1975.

42. Rolfe, J. M.: The secondary task as a measure of mental load. In W. T. Singleton, J. G. Fox, and D. Whitfield (Eds.): Measurement of Man At Work. London, Taylor and Francis Ltd., pp. 135-148, 1971.

43. Roscoe, S. N., and R. C. Williges: Motion Relationships in Aircraft Attitude and Guidance Displays: A Flight Experiment. HUMAN FACTORS, 17:374-387, 1975.

44. Senders, J. W.: Man's Capacity to Use Information From Complex Displays. In H. Quastler (Ed.), Information Theory in Psychology. Glencoe, Illinois, The Free Press, 1955.

45. Senders, J. W.: Tracking With Intermittent Stimuli (forced sampling). Air Research and Development Report No. ARDC-TR-56-8, 1956.

46. Senders, J. W.: Information Input Rates to Human Users: Recent Research Results. WADC Symposium on Air Force Flight Instrumentation Program. Wright-Patterson AFB, Ohio, 1958.

47. Senders, J. W.: The Human Operator as a Monitor and Controller of Multidegree of Freedom Systems. IEEE Transactions on Human Factors in Electronics, Report No. HFE-5(1), pp. 2-5, 1964.

48. Senders, J. W., and K. A. Stevens: Re-analysis of the Pilot Eye-Movement Data. Cambridge, Massachusetts, Bolt, Beranek, and Newman, Inc., Report No. 1136, 1964.

49. Senders, J. W.: The Estimation of Operator Workload in Complex Systems. In K. B. DeGreen (Ed.), Systems Psychology. New York, McGraw-Hill, pp. 207-217, 1970.

50. Spyker, D. A., S. P. Stackhouse, A. S. Khalafalla, and R. C. McLane: Development of Techniques for Measuring Pilot Workload. National Aeronautics and Space Administration Report No. NASA-CR-1888, November 1971.

51. Sternberg, S.: High Speed Scanning in Human Memory. SCIENCE, 153:652-654, 1966.

52. Sternberg, S.: Memory Scanning: New Findings and Current Controversies. QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 27:1-32, 1975.

53. Webster's Third New International Dictionary of the English Language, Unabridged, P. B. Gove (Ed.), Springfield, Massachusetts, G&C Merriam Company, 1967.

54. Weir, D. H., and R. H. Klein: The Measurement and Analysis of Pilot Scanning and Control Behavior During Simulated Instrument Approaches. National Aeronautics and Space Administration Report No. NASA-CR-1535, 1970.

55. White, R. T.: Task Analysis Methods: Review and Development of Techniques for Analyzing Mental Workload in Multiple-Task Situations. McDonnell Douglas Corporation, Report No. MDC J5291, September 1971.

56. Wingert, J. W.: Function Interlace Modifications to Analytic Workload Predictions. In K. D. Gross and J. D. McGrath (Eds.), Crew System Design. Santa Barbara, California, Anacapa Sciences, Inc., 1973.

Bibliography

1. Barber, M. R., C. K. Jones, T. R. Sisk, and F. W. Haise: An Evaluation of the Handling Qualities of Seven General-Aviation Aircraft. National Aeronautics and Space Administration Report No. NASA-TN-D-3726, November 1966.

2. Belcher, J. J.: A Technique for Assessing Operability/Effectiveness of Control-Display Systems. In K. D. Grass and J. J. McGrath (Eds.), Crew System Design. Santa Barbara, California, Anacapa Sciences, Inc., pp. 241-249, July 1973.

3. Bergström, B.: Complex Psychomotor Performance During Different Levels of Experimentally Induced Stress in Pilots. In L. Levi (Ed.), Emotional Stress. New York, American Elsevier Publishing Company, pp. 239-250, 1967.

4. Billings, C. E., R. L. Wick, Jr., R. J. Gerke, and R. C. Chase: The Effects of Alcohol on Pilot Performance During Instrument Flight. FAA Office of Aviation Medicine Report No. FAA-AM-72-4, January 1972.

5. Brictson, C. A.: Operational Measures of Pilot Performance During Final Approach to Carrier Landing. In AGARD Conf. Proc. No. 56, Measurement of Aircrew Performance, the Flight Deck Workload and Its Relation to Pilot Performance, 7-1 to 7-11, December 1969.

6. Chapanis, A., and H. P. Van Cott: Human Engineering Tests and Evaluations. In H. P. Van Cott and R. G. Kinkade (Eds.), Human Engineering Guide to Equipment Design. Washington, D.C., McGraw-Hill Company, pp. 701-728, 1972.

7. Connelly, E. M., F. J. Bourne, D. G. Loental, J. S. Migliaccio, D. A. Burchick, and P. A. Knoop: Candidate T-37 Pilot Performance Measures for Five Contact Maneuvers. Brooks AFB, Texas; Air Force Systems Command Report No. AFHRL-TR-74-88, December 1974.

8. Fraser, T. M.: Philosophy of Simulation in a Man-Machine Space Mission System. National Aeronautics and Space Administration Report No. NASA-SP-102, 1966.

9. Frost, G.: Man-Machine Dynamics. In H. P. Van Cott and R. G. Kinkade (Eds.): Human Engineering Guide to Equipment Design. Washington, D.C., McGraw-Hill Company, pp. 227-310, 1972.

10. Gagne, R. M., and A. W. Melton: Psychological Principles in System Development. New York, Holt, Rinehart, and Winston, 1962.

11. Geiselhart, R., R. I. Koeteeuw, and R. J. Schiffler: A Study of Task Loading Using a Four-Man Crew on a KC-135 Aircraft (giant boom). Wright-Patterson AFB, Ohio, Aeronautical Systems Division Report No. ASD-TR-76-33, April 1977.

12. Hall, E. R., J. F. Parker, Jr., and D. E. Meyer: A Study of Air Force Flight Simulator Programs. Wright-Patterson AFB, Ohio, Aerospace Medical Research Laboratories Report No. AMRL-TR-67-111, June 1967.

13. Hasbrook, A. H., and P. G. Rasmussen: Aural Glide Slope Cues: Their Effect on Pilot Performance During In-flight Simulated ILS Instrument Approaches. FAA Office of Aviation Medicine Report No. FAA-AM-71-24, May 1971.

14. Hasbrook, A. H., and P. G. Rasmussen: In-flight Performance of Civilian Pilots Using Moving-Aircraft and Moving-Horizon Attitude Indicators. FAA Office of Aviation Medicine Report No. FAA-AM-73-9, June 1973.

15. Henry, P. H., J. A. Flueck, J. F. Sanford, H. N. Keiser, R. C. McNee, W. H. Walter III, K. H. Webster, B. O. Hartman, and M. C. Lancaster: Assessment of Performance in a Link GAT-1 Flight Simulator at Three Alcohol Dose Levels. AEROSPACE MEDICINE, 45:33-44, 1974.

16. Hodge, D. C. (Ed.): Standardization of Tasks and Measures for Human Factors Research (Proceedings of a conference held at Texas Tech University). Aberdeen Proving Ground, Maryland, Aberdeen Research and Development Center Report No. TM-19-70, March 1970.

17. Howell, W. C., and I. L. Goldstein: Engineering Psychology, Current Perspectives in Research. New York, Meredith Corporation, 1971.

18. Keenan, J. J., T. C. Parker, and H. P. Lenzycki: Concepts and Practices in the Assessment of Human Performance in Air Force Systems. Wright-Patterson AFB, Ohio, Aerospace Medical Research Laboratories Report No. AMRL-TR-65-168, September 1965.

19. Kidd, J. S.: A New Look at System Research and Analysis. HUMAN FACTORS, 4:209-216, 1962.

20. Kitchin, J. B., and A. Graham: Mental Loading of Process Operators: An Attempt to Device a Method of Analysis and Assessment. ERGONOMICS, 4:1-15, 1961.

21. Knowles, W. B.: Aerospace Simulation and Human Performance Research. HUMAN FACTORS, 9:149-159, 1967.

22. Kraft, C. L., and C. L. Elworth: Flight Deck Workload and Night Visual Approach Performance. In AGARD Conf. Proc. No. 56, Measurement of Aircrew Performance, the Flight Deck Workload and Its Relation to Pilot Performance, 11-1 to 11-14, December 1969.

23. Levison, W. H., J. I. Elkind, and J. L. Ward: Studies of Multivariable Manual Control Systems: A Model for Task Interference. National Aeronautics and Space Administration Report No. NASA-CR-1746, May 1971.

24. Meister, D.: Methods of Predicting Human Reliability in Man-Machine Systems. HUMAN FACTORS, 6:621-646, 1964.

25. Milligan, J. R.: A Student Pilot Automatic Monitoring System. In K. D. Gross and J. J. McGrath (Eds.), Crew System Design. Santa Barbara, California, Anacapa Sciences, Inc., pp. 301-310, July 1973.

26. Parsons, H. M.: Man-Machine System Experiments. Baltimore, The Johns Hopkins Press, 1972.

27. Povenmire, H. K., and S. N. Roscoe: An Evaluation of Ground-Based Flight Trainers in Routine Primary Flight Training. HUMAN FACTORS, 13:109-116, 1971.

28. Rolfe, J. M. (Ed.): Vehicle Simulation for Training and Research. Farnborough Hants, RAF Institute of Aviation Medicine Report No. IAM-442, March 1968.

29. Roscoe, S. N.: Assessment of Pilotage Error in Airborne Area Navigation Procedures. HUMAN FACTORS, 16:223-228, 1974.

30. Schiffler, R. J., R. Geiselhart, and L. Ivey: Crew Composition Study for an Advanced Tanker/Cargo Aircraft (ATCA). Wright-Patterson AFB, Ohio; Deputy for Engineering, Aeronautical Systems Division Report No. ASD-TR-76-20, October 1976.

31. Schohan, B., H. E. Rawson, and S. M. Soliday: Pilot and Observer Performance in Simulated Low Altitude High Speed Flight. HUMAN FACTORS, 7:257-265, 1965.

32. Schultz, W. C., F. D. Newell, and R. F. Whitbeck: A Study of Relationships Between Aircraft System Performance and Pilot Ratings. National Aeronautics and Space Administration Report No. NASA-CR-1643, July 1970.

33. Shackel, B.: Man-Computer Interaction--The Contribution of the Human Sciences. ERGONOMICS, 12:485-499, 1969.

34. Shannon, R. H., and W. L. Waag: Toward the Development of a Criterion for Fleet Effectiveness in the F-4 Fighter Community. Proceedings of the 16th Annual Meeting of the Human Factors Society, pp. 335-340, October 1972.

35. Soliday, S. M., and B. Schohan: Task Loading of Pilots in Simulated Low-Altitude High-Speed Flight. HUMAN FACTORS, 7:45-53, 1965.

36. Wherry, R. J., Jr.: Human Performance Studies for the Airborne Crew Station Design Process. In K. D. Gross and J. J. McGrath (Eds.), Crew System Design. Santa Barbara, California, Anacapa Sciences, Inc., pp. 21-28, July 1973.

37. Zaitzeff, L. P.: Aircrew Task Loading in the Boeing Multimission Simulator. In AGARD Conf. Proc. No. 56, Measurement of Aircrew Performance, the Flight Deck Workload, and Its Relation to Pilot Performance, 8 to 8-3, December 1969.